Regularization and the Bias Variance Tradeoff

Q (Park Effects) Estimate the park effect α of each MLB ballpark, which represents the expected runs scored in one half-inning at that park above that of an average park, if an average offense faces an average defense.

→ Read my full analysis in Appendix of "Grid WAR" paper

<u>training data</u> all half innings from 2017 – 2019

<u>variables</u> i indexes the i^{th} half inning in our dataset
park(i) is the ballpark of half-inning i
ot(i) is the offensive team-szn of half inning i
dt(i) is the defensive team-szn of half-inning i
y_i is the Runs scored in half-inning i

<u>Model</u> $$y_i = \beta_0 + \alpha_{park(i)} + \beta_{ot(i)} + \gamma_{dt(i)} + \varepsilon_i$$

where ε_i is mean zero noise, $\mathbb{E}\,\varepsilon_i = 0$

The park effects α and team quality coefficients β, γ are unknown parameters which need to be estimated from data.

Equivalently, $y_i = x_i^T \beta + \varepsilon_i$

where X is a matrix whose i^{th} Row is defined by

$$
\begin{array}{cccccccccccccc}
 & \text{intercept} & \text{park 1} & \text{park 2} & \cdots & \text{park 30} & \text{ot1} & \text{ot2} & \cdots & \text{ot30} & \text{dt1} & \text{dt2} & \cdots & \text{dt30} \\
x_i^T = & [\,1 & \bullet & \bullet & & \bullet & \bullet & \bullet & & \bullet & \bullet & \bullet & & \bullet\,]
\end{array}
$$

Park block: = 1 at Park(i), 0 else

ot block: 1 at ot(i), 0 else

dt block: 1 at dt(i), 0 else

$\boxed{\text{Problem: Multicollinearity}}$

When home team is on offense, park(i) = ot(i).
When road team is on offense, park(i) = dt(i).
So, it is tough to disentangle $\alpha_{park(i)}$ from $\beta_{ot(i)}$ and $\gamma_{dt(i)}$.

Are the Runs scored in those half-innings due to the offensive home team being good or the park being easy?

To disentangle these effects, we need a huge number of instances of Road teams on offence to figure out $\beta_{ot(i)}$ well,
and a huge number of instances of Home teams on offence to figure out $\gamma_{dt(i)}$ well.
Then, with β_{ot} and γ_{dt} good,
we can figure disentangle α_{park}.

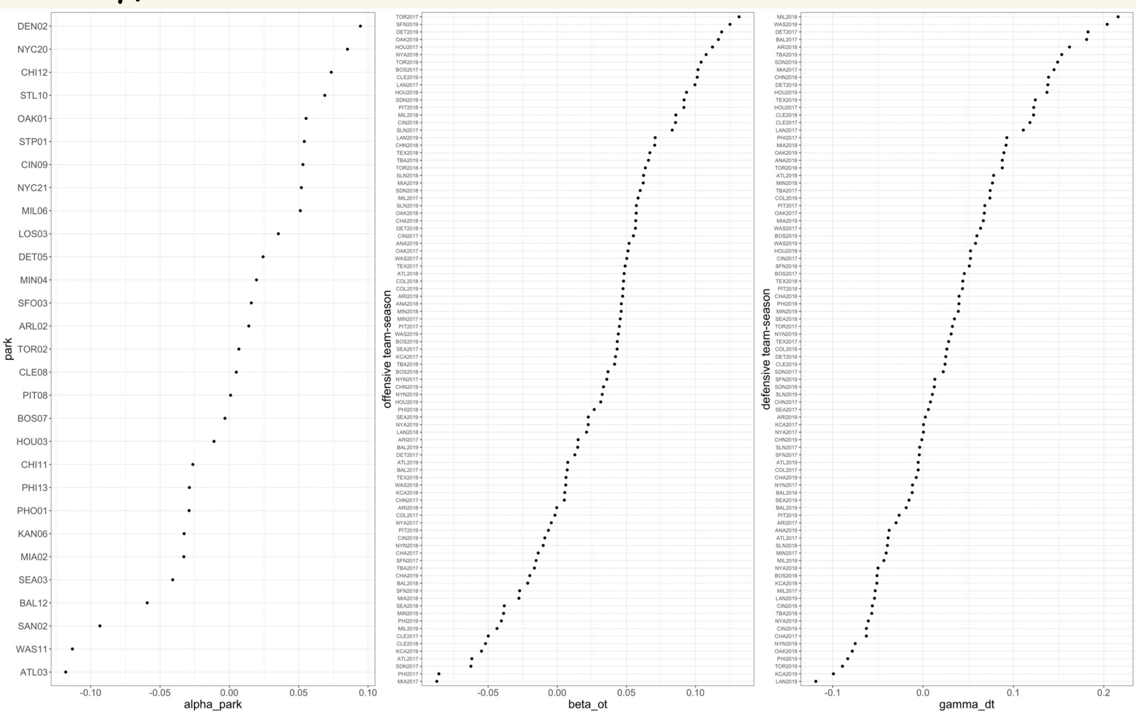
Our current dataset consists of 123,252 half-innings.
This may seem like a lot of data,
but due to our multicollinearity issue
this actually isn't a huge amount of data.
To demonstrate, we run a simulation study.

Q How much does multicollinearity affect our park effect estimates?
How well does OLS recover the park effects?

Simulation Study

Idea Pretend we knew the true coefficients, generate simulated data, and see how well we estimate the coefficients.

* Suppose the true coefficients are



which are chosen to have a "Reasonable" scale.

* Then, assuming our model is <u>true</u>, let's generate the Response y vector (Runs scored in a half inning) M times according to

$$\boxed{y_i = \text{Round}\left(\mathcal{N}_+\left(x_i^T\beta, 1\right)\right)}$$

where x_i^T is the i^{th} half inning from our observed data matrix of all 123,252 half-innings from 2017 to 2019.

example snippet of simulated y

```
      [,1]
 [1,]    1
 [2,]    1
 [3,]    1
 [4,]    2
 [5,]    1
 [6,]    2
 [7,]    1
 [8,]    0
 [9,]    1
[10,]    1
```

Here,

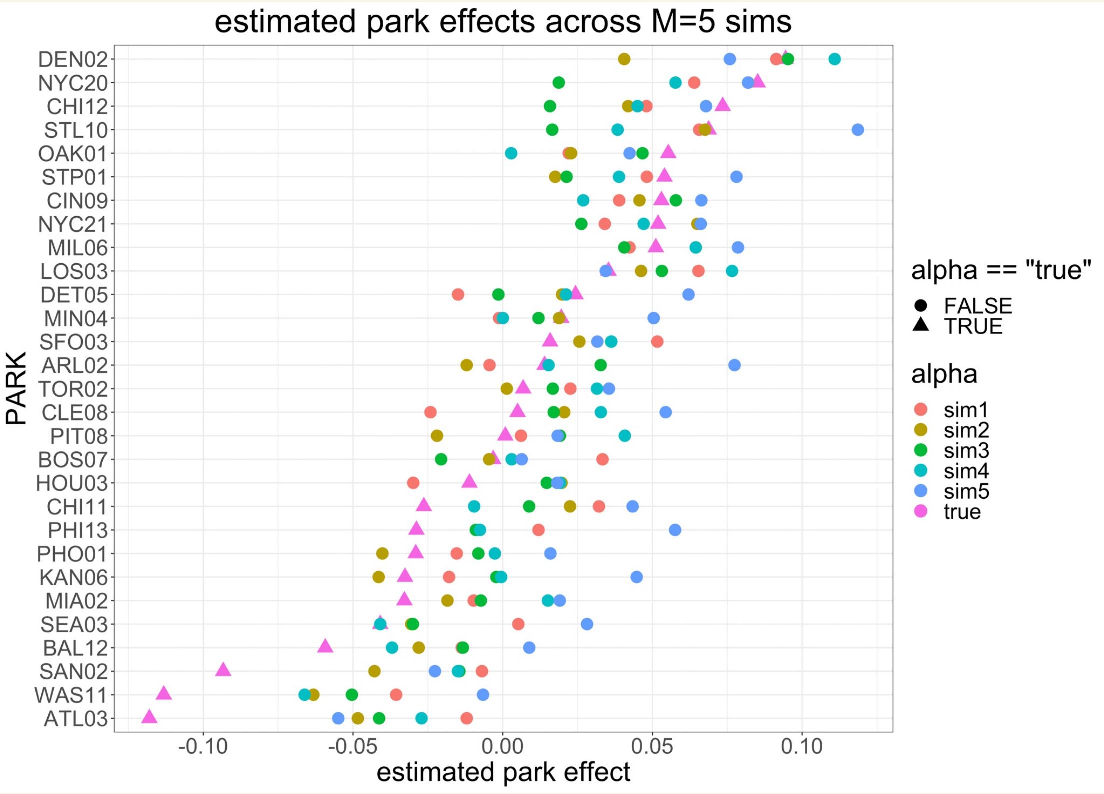
- $\mathcal{N}_+$ means normal dist. conditional on it being ≥ 0
- "Round" because Runs scored is an integer ≥ 0
- $\mathbb{E}\, y_i \approx x_i^T\beta$

  equivalently, $y_i \approx x_i^T\beta + \varepsilon_i$, $\mathbb{E}\,\varepsilon_i = 0$

  so our original model assumption holds true even if we don't explicitly write ε_i here

* Then, let's use linear regression to estimate the coefficients $\hat{\beta}$ on each of our M simulated datasets (X, y) and see how well we recover the park effects! $\hat{\beta} = (X^T X)^{-1} X^T y$.

We can do this because it's a simulation and we know the "true" park effects.



* Due to <u>Randomness</u> in the training dataset, from the noise in generating y, each simulation yields very <u>different</u> park effects estimates $\hat{\alpha}$, even though the "true" park effects are the same.

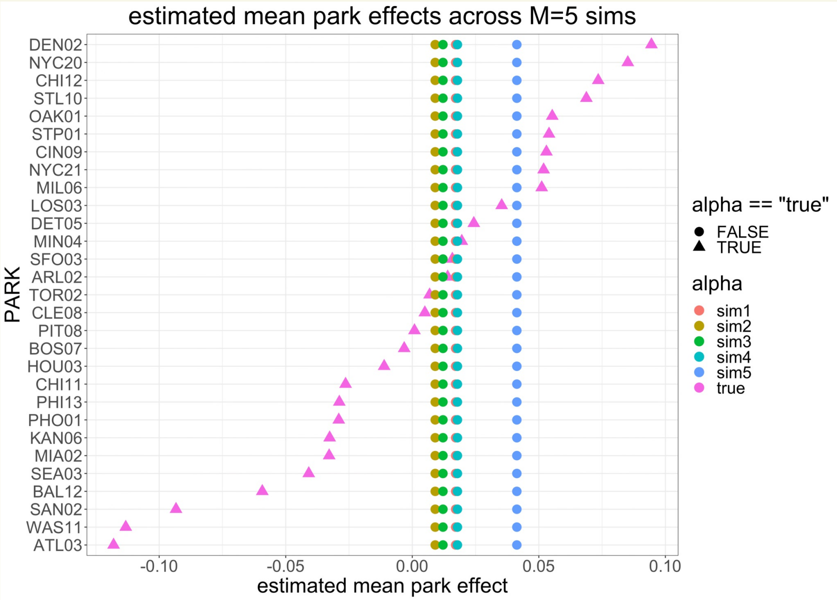
* The OLS (ordinary least squares = ordinary linear regression) coefficients $\hat{\alpha}^{(OLS)}$ change quite significantly across different simulations; they are quite sensitive to the noise of the training set

* How can we make the coefficients <u>less sensitive</u> to the random idiosyncracies of our training set?

Q What's the least sensitive estimator you can think of?

overall mean $\bar{\hat{\alpha}}$ estimated mean park effect

zero the constant value 0



* Constant values like zero, or overall mean — not too sensitive to the random idiosyncracies of the training set, but are wrong for many parks
* OLS park effect estimates — very sensitive to the randomness of the training set, but are unbiased (on average, i.e. averaged over many training set generations, they are in the right spot)

Q How can we blend the strengths of OLS with the strengths of the overall mean?

Idea Shrink the OLS estimates towards a constant value, like the overall mean or to zero.

The latter is easier, so let's go with that.

In other words:

> **Idea** Shrink the OLS estimates towards zero, i.e. just make them smaller, which will make them less sensitive!

* In ordinary linear regression, we estimate the coefficients β by minimizing the Residual sum of squares,

$$\hat{\beta}^{(OLS)} = \underset{\beta}{\operatorname{argmin}} \; RSS(\beta)$$

$$= \underset{\beta}{\operatorname{argmin}} \; \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

* In Ridge Regression we instead minimize the RSS with a penalty term that encourages the estimated coefficients $\hat{\beta}$ to be smaller (i.e., to lie closer to 0),

$$\hat{\beta}^{(Ridge)} = \underset{\beta}{\operatorname{argmin}} \; \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j} \beta_j^2$$

* This technique of adding a penalty term to the loss function we are minimizing is called Regularization.

The hyperparameter $\lambda > 0$ describes by how much we are penalized for having large β_j.

λ is simply a number, which is tuned using cross-validation.

Large λ $\rightarrow$ large penalty for large β_j
$\rightarrow$ forces β_j to be smaller.

$\lambda = 0$ $\rightarrow$ equivalent to OLS
$\rightarrow$ no shrinkage of β.

$$\beta^{(Ridge)} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - X_i^T \beta)^2 + \lambda \sum_j \beta_j^2$$

$$= \underset{\beta}{\operatorname{argmin}} \, (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

in matrix notation.

Calulus: Set gradient equal to 0 and solve!

$$L(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

$$= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta + \lambda \beta^T \beta$$

$$\nabla_\beta L(\beta) = -2X^T y - 2X^T X \beta + 2\lambda\beta = 0$$

$$\Longrightarrow (X^T X + \lambda I)\beta = X^T y$$

$$\Longrightarrow \boxed{\hat{\beta}^{(ridge)} = (X^T X + \lambda I)^{-1} X^T y}$$

Solution always exists when $\lambda > 0$.

<u>Ridge</u> Regression — add matrix

$$\lambda I = \begin{pmatrix} \lambda & & 0 \\ & \lambda & \\ 0 & & \ddots \ \lambda \end{pmatrix} \text{ to } X^T X$$

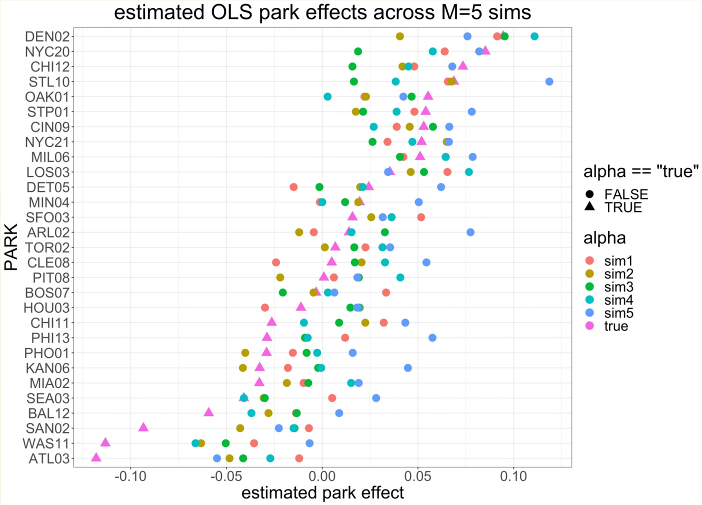
prior to inverting. This is a "ridge" of λ's.

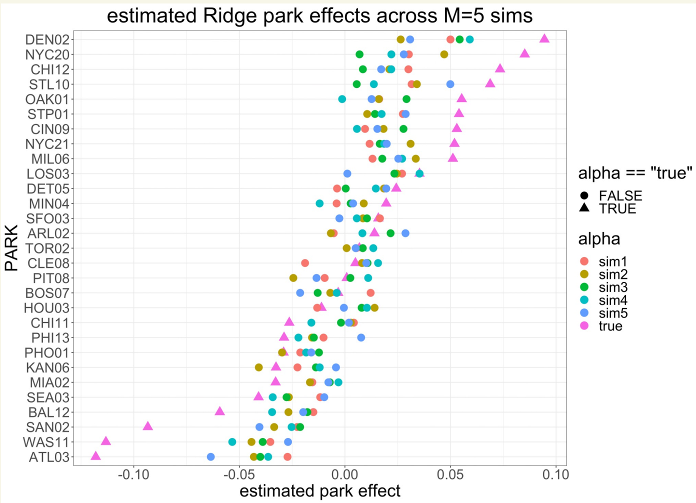
$(X^T X + \lambda I)^{-1}$ is like multiplying by $\dfrac{1}{\bullet + \lambda}$,

$(X^T X)^{-1}$ is like multiplying by $\dfrac{1}{\bullet}$

adding $\lambda > 0$ to the denominator

<u>shrinks</u> the estimates $\hat{\beta}$!



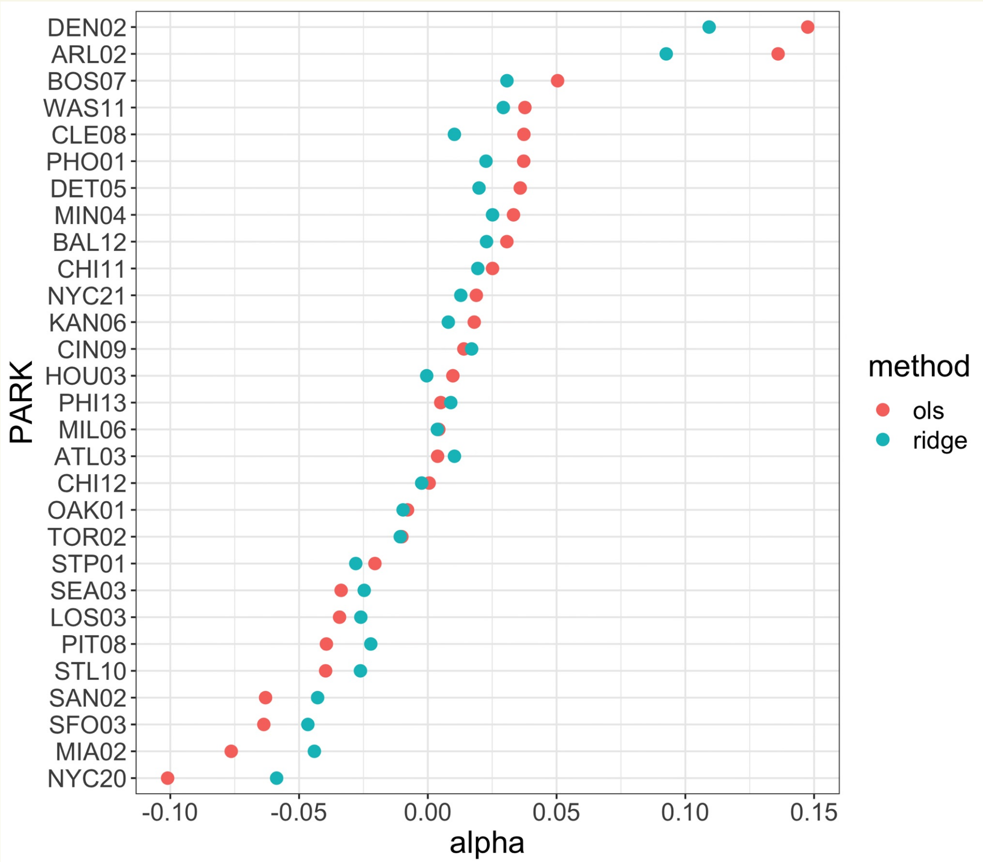


* Ridge regression park effect estimates indeed are more stable across simulations, i.e. are less sensitive to the noise of the training set!

```
> ### error
> err(beta.pk.df.sim)
[1] 0.03528335
> err(beta.pk.df.sim_ridge)
[1] 0.03804942
> ### error on non-outliers
> err(beta.pk.df.sim %>% filter( abs(beta.pk.true) < 0.05) )
[1] 0.02533202
> err(beta.pk.df.sim_ridge %>% filter( abs(beta.pk.true) < 0.05) )
[1] 0.01690153
> ### error on outliers
> err(beta.pk.df.sim %>% filter( abs(beta.pk.true) >= 0.05) )
[1] 0.04406852
> err(beta.pk.df.sim_ridge %>% filter( abs(beta.pk.true) >= 0.05) )
[1] 0.05359246
```

* Shrinking outliers isn't always a great idea; OLS outperforms on outliers

* Park effects on Real MLB data, 2017-2019



* We see that Ridge indeed shrinks the park effects towards zero!
* On this Real data, it turns out that the Ridge shrunken park effects are everywhere better than OLS since OLS overfits...
(based on out-of-sample predictive performance)

Q How do we quantify the sensitivity of an estimator to the random idiosyncrasies of a training dataset?

<u>Model</u> Suppose $y_i = f(x_i) + \varepsilon_i$

for some "true" underlying function f

and noise ε_i with $\mathbb{E}\varepsilon_i = 0$.

<u>Goal</u> is to estimate f with $\hat{f}$

e.g.
- OLS
- Ridge
- overall mean

etc.

training dataset $D = \{(x_i, y_i)\}_{i=1}^n$

$$\hat{f} = \hat{f}(x; D)$$

<u>Want</u> our estimator $\hat{f}$ to be as "close" to true f as possible, which we can measure from data as the smallest out-of-sample Mean Squared Error which uses datapoints (x, y) not in the training dataset,

$$MSE(f, \hat{f}) := \mathbb{E}\left[Y(x) - \hat{f}(x; D)\right].$$

$$\mathrm{MSE}(x;D) = \mathbb{E}\left(Y - \hat{f}(x;D)\right)^2$$

$$= \mathbb{E}(Y - \hat{f})^2$$

using $\hat{f} = \hat{f}(x;D)$ as shorthand

$$= \mathbb{E}\left(Y^2 - 2Y\hat{f} + \hat{f}^2\right)$$

$$= \mathbb{E}Y^2 - 2\mathbb{E}(Y\hat{f}) + \mathbb{E}\hat{f}^2$$

$$= \mathbb{E}(f(x)+\varepsilon)^2 - 2\mathbb{E}\left[(f(x)+\varepsilon)\hat{f}\right] + \mathbb{E}\hat{f}^2$$

since $Y = f(x) + \varepsilon$

$$= \mathbb{E}\left(f^2 + 2f\varepsilon + \varepsilon^2\right) - 2\mathbb{E}\left(f\hat{f} + \hat{f}\varepsilon\right) + \mathbb{E}\hat{f}^2$$

using $f = f(x)$ as shorthand

$$= f^2 + 2f\cancelto{0}{\mathbb{E}\varepsilon} + \mathbb{E}\varepsilon^2 - 2f\mathbb{E}\hat{f}$$
$$-2\mathbb{E}\hat{f}\,\cancelto{0}{\mathbb{E}\varepsilon} + \mathbb{E}\hat{f}^2$$

since $f(x)$ is deterministic and not random
and $\hat{f}(x)$ is independent of ε

$$= f^2 - 2f\mathbb{E}\hat{f} + \mathbb{E}\hat{f}^2 + \mathbb{E}\varepsilon^2$$

$$= f^2 - 2f\mathbb{E}\hat{f} + \underline{(\mathbb{E}\hat{f})^2} + \mathbb{E}\hat{f}^2 \underline{- (\mathbb{E}\hat{f})^2} + \mathbb{E}\varepsilon^2$$

$$= (f - \mathbb{E}\hat{f})^2 + [\mathbb{E}\hat{f}^2 - (\mathbb{E}\hat{f})^2] + \mathbb{E}\varepsilon^2$$

$$\doteq \text{Bias}(\hat{f})^2 + \text{Var}(\hat{f}) + \sigma_\varepsilon^2.$$

$$\boxed{\begin{array}{l}\text{Bias Variance Tradeoff} \\ MSE(x;D) = [\text{Bias}\ \hat{f}(x;D)]^2 + \text{Var}[\hat{f}(x;D)] + \sigma_\varepsilon^2\end{array}}$$

* $\sigma_\varepsilon^2 = \mathbb{E}\varepsilon^2$ is irreducible error, the noise inherent to the problem (e.g. for park effects, σ_ε^2 is the inherent noise of scoring a certain number of runs in a half inning)

* $\text{Var}(\hat{f}) = \mathbb{E}\hat{f}^2 - (\mathbb{E}\hat{f})^2$ is how variable $\hat{f}$ is depending on the training set, i.e. how much it responds to randomness in the training set

* $\text{Bias}(\hat{f})^2 = (f - \mathbb{E}\hat{f})^2$
is how close $\hat{f}$ is to f
on average (averaged over $N \uparrow \infty$
training sets of the same size)

* Bias-Variance Tradeoff for Park Effects:

1. Overall mean — very low variance, very high bias
2. OLS — low bias, high variance
3. Ridge — introduces bias with the penalty term $+\lambda \sum_j \beta_j^2$ in order to lower variance!

Takeaway: To make better predictions, sometimes it helps to introduce some bias to lower the variance!